Estimating return levels from serially dependent extremes

Lee Fawcett and Dave Walshaw

Newcastle University, Newcastle upon Tyne, U.K.

Sheffield University, May 17th 2012

Structure of this talk

1. Background

- Motivating examples
- The data: wind speeds and sea–surges
- Statistical modelling of extremes

2. Simulation study

- The data
- Return levels: dependence on serial correlation
- The extremal index
- Some results

3. Data applications

- Newlyn sea–surges
- Bradfield wind speeds
- Bootstrapping for confidence intervals

1.1 Background and motivation: extreme wind speeds

In the U.K., the *British Standards Institution* produce contour maps displaying strength requirements for structures based on “once–in–50–year gust speeds”.

This is known as the **50–year return level** gust.

The maps themselves are the result of simple extreme value analyses carried out on medium to long term records collected at stations in the U.K.

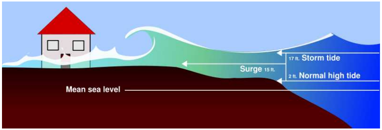
1.1 Background and motivation: extreme wind speeds

During storms in 1987, 2002 and 2005, gust speeds exceeded the **200–year return level**.

- Perhaps building codes should be revised?
- Or maybe the estimation procedure is inappropriate...

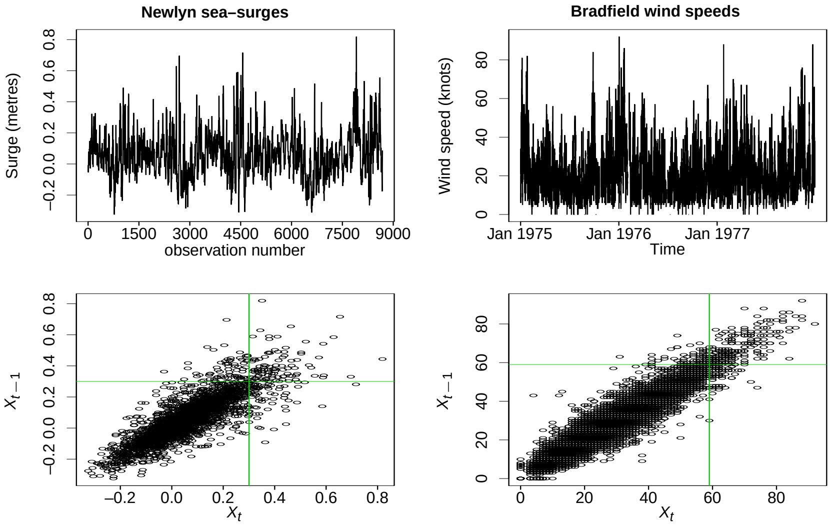
1.1 Background and motivation: extreme sea–surge

Sea level = mean sea level + tide + surge + waves



- Sea surge is generated by wind and air pressure
- Key factor in coastal flooding — e.g. North Sea (1953), Bangladesh (1993)
- Practical motivation: **structural failure** — probably a sea–wall in this case — is likely under the condition of extreme surges
- Aim: Design a sea–wall so that it protects against the once–in–a–hundred year sea surge, or the **100–year return level**

1.2 The data: Bradfield gusts and Newlyn sea–surges



1.3 Modelling extremes

Let $\{X_n\}$ denote a stationary sequence of random variables with common distribution function F, and let $M_n = \max\{X_1, \ldots, X_n\}$.

It is typically the case that, as $n \to \infty$,

$$\Pr(M_n \leq x) \quad \approx \quad F^{n\theta}(x), \tag{1}$$

where $\theta \in (0, 1)$ is known as the **extremal index**; see, for e.g., Leadbetter & Rootzén (1988).

As $\theta \to 0$ there is increasing dependence in the extremes of the process; for an independent process, $\theta = 1$.

1.3 Modelling extremes: The GPD

The **Generalised Extreme Value** distribution (GEV) is the limiting model for F^n.

Pickands (1975) showed that for large enough u, $(X - u | X > u)$ follows a **Generalised Pareto distribution** (GPD) with distribution function

$$\mathcal{H}(y; \sigma, \xi) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}, \qquad (2)$$

defined on $\{y : y > 0 \text{ and } (1 + \xi y/\sigma) > 0\}$, where $\sigma > 0$ and ξ are the GPD scale and shape (respectively).

- The GPD provides a natural way of modelling extremes of time series such as sea–surge/wind speed extremes
- Much less wasteful than the standard “annual maxima” approach using the GEV
- What about serial dependence? It is usually the case that $\theta < 1$ in Equation (1): **Peaks Over Thresholds** (POT) with declustering interval κ
- Other issues, e.g. seasonal variability: Fourier forms for the GPD parameters, piecewise seasonality approach,...

1.3 Modelling extremes: return levels

Suppose the GPD is a suitable model for threshold exceedances $(X - u)$. Then

$$\begin{aligned}\Pr(X > x | X > u) &= \frac{\Pr(X > x \cap X > u)}{\Pr(X > u)} \\ &= \frac{\Pr(X > x)}{\Pr(X > u)},\end{aligned}$$

for $x > u$. This leads to

$$\Pr(X \leq x) = 1 - \lambda_u \left[1 + \xi \left(\frac{x - u}{\sigma}\right)\right]^{-1/\xi}, \qquad (3)$$

where $\lambda_u = \Pr(X > u)$.

1.3 Modelling extremes: return levels

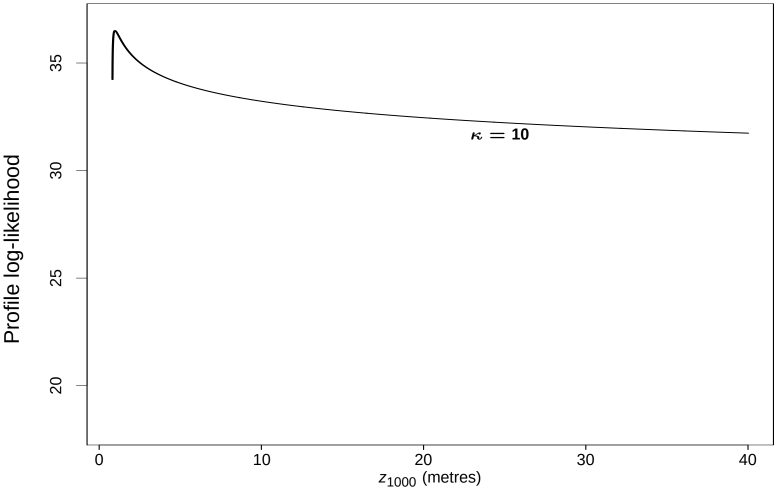
Estimates of an extreme quantile z_s can then be obtained by equating (1) to $1 - s^{-1}$, where $F^n(x)$ is given by (3), and solving for $x = z_s$.

- z_s is the ***s*–observation return level** associated with **return period** s
- We usually work on an annual scale, giving the r*–year* return level

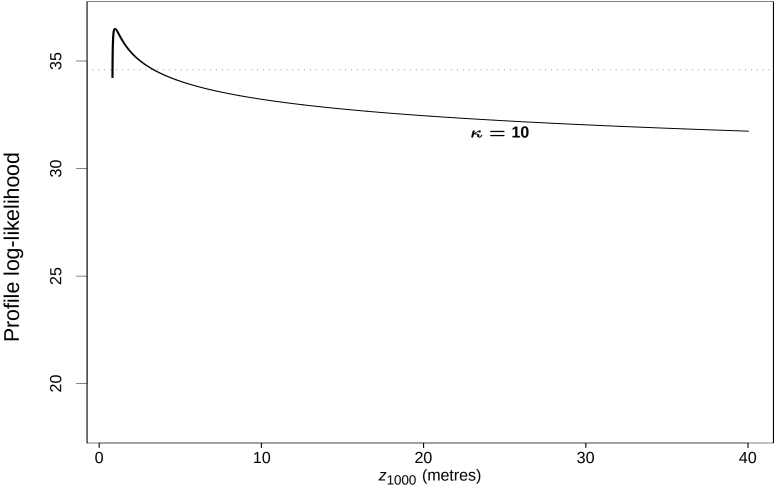
$$z_r = u + \frac{\sigma}{\xi}\left[\left(\lambda_u^{-1}\left\{1 - [1 - 1/(rn_y)]^{\theta^{-1}}\right\}\right)^{-\xi} - 1\right] \quad (4)$$

  where n_y is the number of observations per year.
- In practice, $\hat{z}_r$ is often obtained by replacing (λ, σ, ξ) in (4) with MLEs $(\hat{\lambda}_u, \hat{\sigma}, \hat{\xi})$; working with cluster peaks from a POT analysis, $\theta \approx 1$.
- Confidence intervals are usually constructed using profile likelihood.

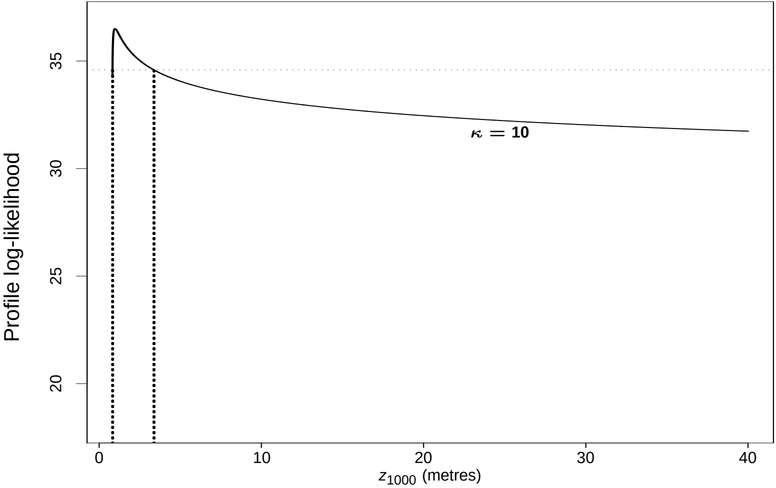
1.4 Return levels: Newlyn sea surges



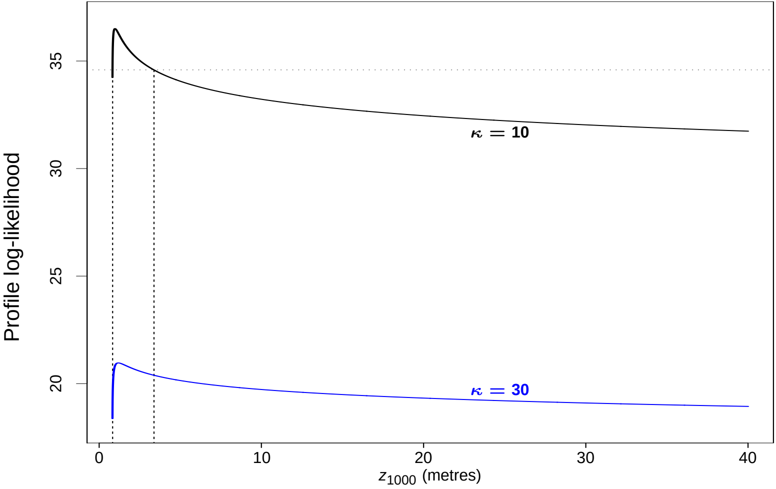
1.4 Return levels: Newlyn sea surges



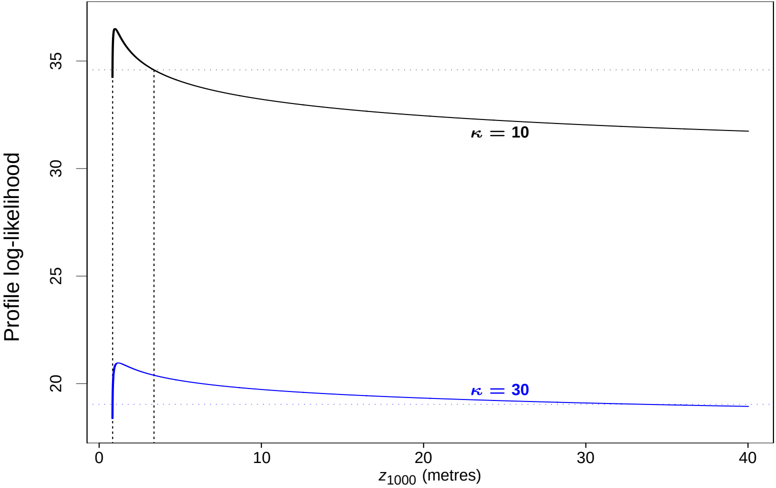
1.4 Return levels: Newlyn sea surges



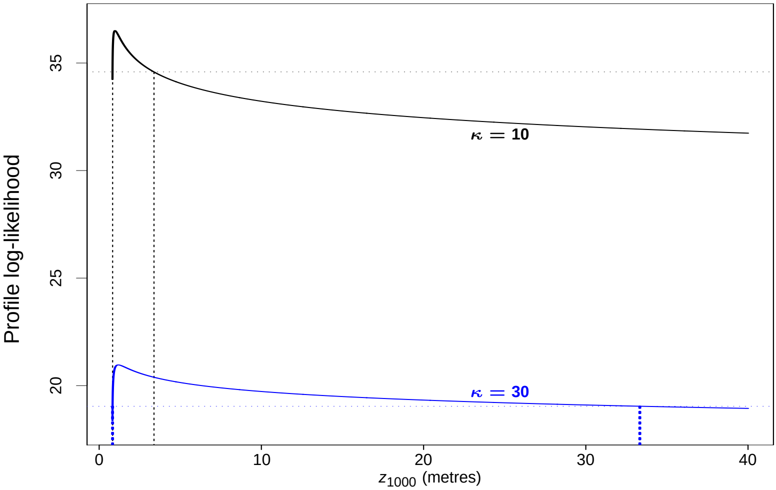
1.4 Return levels: Newlyn sea surges



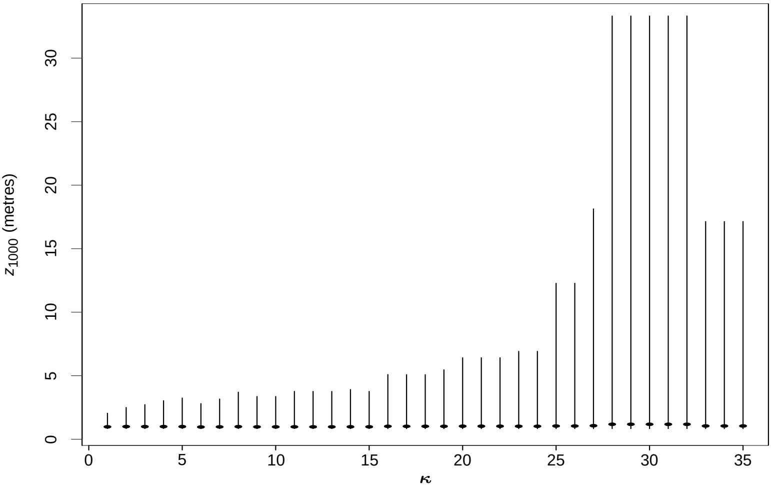
1.4 Return levels: Newlyn sea surges



1.4 Return levels: Newlyn sea surges



1.4 Return levels: Newlyn sea surges



2 Simulation study: Aims

- To investigate the use of *all* threshold excesses for estimating return levels over the standard POT approach
- Will need to consider the issue of serial dependence carefully
- This will require consideration of the **extremal index**
- If successful, we will
  - Avoid declustering altogether
  - Press more extremes into use $\longrightarrow$ increase estimation precision of return levels

2.1 Simulated data

Simulate Markov chains with joint density given by

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n-1} f(x_i, x_{i+1}; \psi) \bigg/ \prod_{i=2}^{n-1} f(x_i; \phi), \qquad i = 1, \ldots, n-1.$$

- GPD density used for contributions to the denominator
- Invoke bivariate extreme value theory for contributions to the numerator on $(u, \infty) \times (u, \infty)$

2.1 Simulated data

We make use of two well–known *symmetric* dependence models for the generation of consecutive extremes:

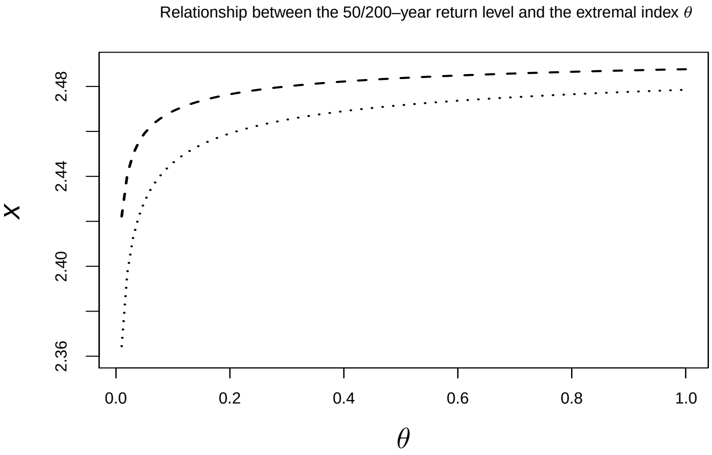
- The **logistic** model, with dependence parameter α, $0 < \alpha \leq 1$
- The **negative logistic** model, with dependence parameter $\rho > 0$
- Independence: $\alpha = 1$ or $\rho \searrow 0$
- Complete dependence: $\alpha \searrow 0$ or $\rho \to \infty$.

2.1 Simulated data

We also use a model allowing for *asymmetry* in the dependence structure:

- The **bilogistic** model, with dependence parameters (α, β)
- $\alpha - \beta$ determines the extent of asymmetry
- Independence: $\alpha = \beta \to 1$
- Reduces to the logistic model when $\alpha = \beta$

2.2 Return levels: dependence on serial correlation



2.2 Connection with dependence models

Define, arbitrarily, x_m such that $F^n(x_m) = 1/2$. Then from (1):

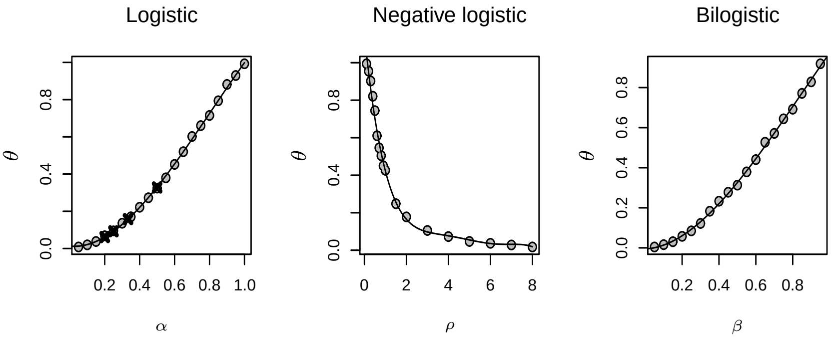
$$\Pr(\max\{X_1, \ldots, X_m\} \leq x_m)^{1/\theta_m} \approx \frac{1}{2}, \qquad \text{giving}$$

$$\theta_m \approx -\frac{\log \Pr(\max\{X_1, \ldots, X_m\} \leq x_m)}{\log 2}.$$

This provides a link between the dependence parameter(s) for any model for extremal dependence and the extremal index. We can

- simulate M first–order Markov chains each of length m with logistic(α)/negative logistic(ρ)/bilogistic(α, β) dependence
- estimate the numerator in the above as the proportion of simulated chains whose maximum does not exceed x_m

2.2 Connection with dependence models



2.3 Extremal index estimation

- A **polynomial estimator** ($\hat{\theta}^{[1]}_{\text{log}}$, $\hat{\theta}^{[1]}_{\text{neglog}}$ and $\hat{\theta}^{[1]}_{\text{bilog}}$)
  - Fit one of the dependence models (logistic/negative logistic/bilogistic) to consecutive pairs and estimate the dependence parameter(s) in that model
  - Use the polynomial relationships previously obtained to estimate θ

- Other commonly–used estimators
  - **Cluster size estimators** ($\hat{\theta}^{[2]}$, $\hat{\theta}^{[3]}$)
  - A **maxima method** ($\hat{\theta}^{[4]}$)
  - An **intervals estimator** ($\hat{\theta}^{[5]}$)

2.4 Simulation study details

Simulate stationary first–order Markov chains of extreme value type according to the three models given previously. We use

- $\alpha = 0.10, 0.11, \ldots, 1$
- $\rho = 0.10, 0.15, \ldots, 1, 1.1, \ldots, 7.0,$
- $\alpha = 0.6$ and $\beta = 0.10, 0.11, \ldots, 0.99$

for the **logistic**, **negative logistic** and **bilogistic** (respectively).

Also simulated data with non–extremal dependence – $AR(1)$ process.

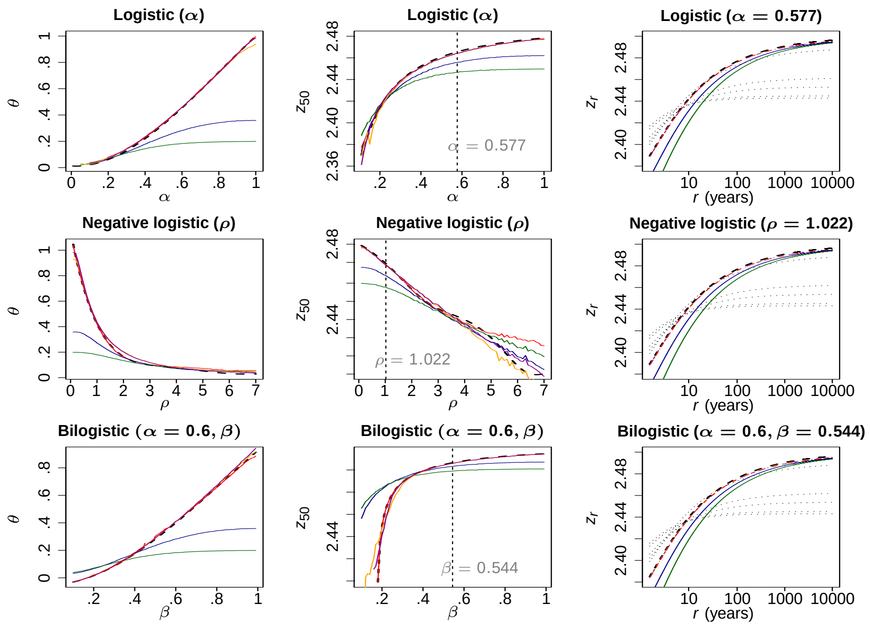
The marginals are transformed to GPD(λ_u, σ, ξ), using $\lambda_u \approx 0.05$, $\sigma = 1$ and $\xi = -0.4, -0.1, 0, 0.3, 0.8$.

2.4 Simulation study details

We simulate $N = 1000$ chains of length $n = 10,000$; for each, we fit the GPD to:

- *all* excesses over u, giving $\left(\hat{\lambda}_u, \hat{\sigma}, \hat{\xi}, \hat{\theta}^{[1]}, \ldots, \hat{\theta}^{[5]}\right)^{(j)} \longrightarrow \hat{z}_r^{(j)}$, $j = 1, \ldots, 1000$
- *cluster peak* excesses over u, using $\kappa = 5, 20, 30, 50, 60$, giving $\left(\hat{\lambda}_u, \hat{\sigma}, \hat{\xi}\right)^{(j)} \longrightarrow \hat{z}_r^{(j)}$, $j = 1, \ldots, 1000$

2.5 Some results



3.1 Data applications: Newlyn sea surges

	$\hat{\theta}$	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
All excesses				
using $\hat{\theta}^{[1]}_{\text{log}}$	0.425 (0.045)	0.817 (0.073)	0.903 (0.107)	1.034 (0.179)
using $\hat{\theta}^{[1]}_{\text{neglog}}$	0.413 (0.037)	0.816 (0.073)	0.902 (0.107)	1.033 (0.178)
using $\hat{\theta}^{[1]}_{\text{bilog}}$	0.377 (0.020)	0.810 (0.071)	0.897 (0.105)	1.029 (0.176)
using $\hat{\theta}^{[2]}$	0.182 (0.047)	0.767 (0.059)	0.860 (0.090)	1.000 (0.159)
using $\hat{\theta}^{[3]}$	0.106 (0.032)	0.732 (0.052)	0.830 (0.079)	0.978 (0.146)
using $\hat{\theta}^{[4]}$	0.282 (0.206)	0.793 (0.078)	0.883 (0.105)	1.024 (0.171)
using $\hat{\theta}^{[5]}$	0.223 (0.050)	0.779 (0.062)	0.870 (0.094)	1.018 (0.163)
Cluster peaks	—	0.868 (0.106)	0.920 (0.144)	0.975 (0.202)

Maximum likelihood estimates for the extremal index and three return levels for the Newlyn sea–surges (units for return levels are in metres).

- Suitability of first order Markov assumption/models used?
- Dependence of some estimators for θ on auxiliary parameter
- $\hat{\theta}^{[5]}$ most suitable?

3.1 Data applications: Newlyn sea surges

3.1 Data applications: Newlyn sea surges

	$\hat{\theta}$	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
All excesses				
using $\hat{\theta}^{[1]}_{\text{log}}$	0.425 **(0.045)**	0.817 **(0.073)**	0.903 **(0.107)**	1.034 **(0.179)**
using $\hat{\theta}^{[1]}_{\text{neglog}}$	0.413 **(0.037)**	0.816 **(0.073)**	0.902 **(0.107)**	1.033 (0.178)
using $\hat{\theta}^{[1]}_{\text{bilog}}$	0.377 **(0.020)**	0.810 **(0.071)**	0.897 **(0.105)**	1.029 (0.176)
using $\hat{\theta}^{[2]}$	0.182 (0.047)	0.767 (0.059)	0.860 (0.090)	1.000 (0.159)
using $\hat{\theta}^{[3]}$	0.106 (0.032)	0.732 (0.052)	0.830 (0.079)	0.978 (0.146)
using $\hat{\theta}^{[4]}$	0.282 (0.206)	0.793 (0.078)	0.883 (0.105)	1.024 (0.171)
using $\hat{\theta}^{[5]}$	0.223 (0.050)	0.779 (0.062)	0.870 (0.094)	1.018 (0.163)
Cluster peaks	—	0.868 **(0.106)**	0.920 **(0.144)**	0.975 **(0.202)**

Maximum likelihood estimates for the extremal index and three return levels for the Newlyn sea–surges (units for return levels are in metres).

- Suitability of first order Markov assumption/models used?
- Dependence of some estimators for θ on auxiliary parameter
- $\hat{\theta}^{[5]}$ most suitable?

3.1 Data applications: Newlyn sea surges

	$\hat{\theta}$	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
All excesses				
using $\hat{\theta}^{[1]}_{\text{log}}$	0.425 (0.045)	0.817 (0.073)	0.903 (0.107)	1.034 (0.179)
using $\hat{\theta}^{[1]}_{\text{neglog}}$	0.413 (0.037)	0.816 (0.073)	0.902 (0.107)	1.033 (0.178)
using $\hat{\theta}^{[1]}_{\text{bilog}}$	0.377 (0.020)	0.810 (0.071)	0.897 (0.105)	1.029 (0.176)
using $\hat{\theta}^{[2]}$	0.182 (0.047)	0.767 (0.059)	0.860 (0.090)	1.000 (0.159)
using $\hat{\theta}^{[3]}$	0.106 (0.032)	0.732 (0.052)	0.830 (0.079)	0.978 (0.146)
using $\hat{\theta}^{[4]}$	0.282 (0.206)	0.793 (0.078)	0.883 (0.105)	1.024 (0.171)
using $\hat{\theta}^{[5]}$	0.223 (0.050)	0.779 (0.062)	0.870 (0.094)	1.018 (0.163)
Cluster peaks	—	0.868 (0.106)	0.920 (0.144)	0.975 (0.202)

Maximum likelihood estimates for the extremal index and three return levels for the Newlyn sea–surges (units for return levels are in metres).

- Suitability of first order Markov assumption/models used?
- Dependence of some estimators for θ on auxiliary parameter
- $\hat{\theta}^{[5]}$ most suitable?

3.1 Data applications: Newlyn sea surges

	$\hat{\theta}$	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
All excesses				
using $\hat{\theta}^{[1]}_{\text{log}}$	0.425 (0.045)	0.817 (0.073)	0.903 (0.107)	1.034 (0.179)
using $\hat{\theta}^{[1]}_{\text{neglog}}$	0.413 (0.037)	0.816 (0.073)	0.902 (0.107)	1.033 (0.178)
using $\hat{\theta}^{[1]}_{\text{bilog}}$	0.377 (0.020)	0.810 (0.071)	0.897 (0.105)	1.029 (0.176)
using $\hat{\theta}^{[2]}$	0.182 **(0.047)**	0.767 **(0.059)**	0.860 **(0.090)**	1.000 **(0.159)**
using $\hat{\theta}^{[3]}$	0.106 **(0.032)**	0.732 **(0.052)**	0.830 **(0.079)**	0.978 **(0.146)**
using $\hat{\theta}^{[4]}$	0.282 **(0.206)**	0.793 **(0.078)**	0.883 **(0.105)**	1.024 **(0.171)**
using $\hat{\theta}^{[5]}$	0.223 (0.050)	0.779 (0.062)	0.870 (0.094)	1.018 (0.163)
Cluster peaks	—	0.868 **(0.106)**	0.920 **(0.144)**	0.975 **(0.202)**

Maximum likelihood estimates for the extremal index and three return levels for the Newlyn sea–surges (units for return levels are in metres).

- Suitability of first order Markov assumption/models used?
- Dependence of some estimators for θ on auxiliary parameter
- $\hat{\theta}^{[5]}$ most suitable?

3.1 Data applications: Newlyn sea surges

	$\hat{\theta}$	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
All excesses				
using $\hat{\theta}^{[1]}_{\text{log}}$	0.425 (0.045)	0.817 (0.073)	0.903 (0.107)	1.034 (0.179)
using $\hat{\theta}^{[1]}_{\text{neglog}}$	0.413 (0.037)	0.816 (0.073)	0.902 (0.107)	1.033 (0.178)
using $\hat{\theta}^{[1]}_{\text{bilog}}$	0.377 (0.020)	0.810 (0.071)	0.897 (0.105)	1.029 (0.176)
using $\hat{\theta}^{[2]}$	0.182 (0.047)	0.767 (0.059)	0.860 (0.090)	1.000 (0.159)
using $\hat{\theta}^{[3]}$	0.106 (0.032)	0.732 (0.052)	0.830 (0.079)	0.978 (0.146)
using $\hat{\theta}^{[4]}$	0.282 (0.206)	0.793 (0.078)	0.883 (0.105)	1.024 (0.171)
using $\hat{\theta}^{[5]}$	0.223 (0.050)	0.779 (0.062)	0.870 (0.094)	1.018 (0.163)
Cluster peaks	—	0.868 (0.106)	0.920 (0.144)	0.975 (0.202)

Maximum likelihood estimates for the extremal index and three return levels for the Newlyn sea–surges (units for return levels are in metres).

- Suitability of first order Markov assumption/models used?
- Dependence of some estimators for θ on auxiliary parameter
- $\hat{\theta}^{[5]}$ most suitable?

3.1 Data applications: Newlyn sea surges

	$\hat{\theta}$	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
All excesses				
using $\hat{\theta}^{[1]}_{\text{log}}$	0.425 (0.045)	0.817 (0.073)	0.903 (0.107)	1.034 (0.179)
using $\hat{\theta}^{[1]}_{\text{neglog}}$	0.413 (0.037)	0.816 (0.073)	0.902 (0.107)	1.033 (0.178)
using $\hat{\theta}^{[1]}_{\text{bilog}}$	0.377 (0.020)	0.810 (0.071)	0.897 (0.105)	1.029 (0.176)
using $\hat{\theta}^{[2]}$	0.182 (0.047)	0.767 (0.059)	0.860 (0.090)	1.000 (0.159)
using $\hat{\theta}^{[3]}$	0.106 (0.032)	0.732 (0.052)	0.830 (0.079)	0.978 (0.146)
using $\hat{\theta}^{[4]}$	0.282 (0.206)	0.793 (0.078)	0.883 (0.105)	1.024 (0.171)
using $\hat{\theta}^{[5]}$	0.223 **(0.050)**	0.779 **(0.062)**	0.870 **(0.094)**	1.018 **(0.163)**
Cluster peaks	—	0.868 **(0.106)**	0.920 **(0.144)**	0.975 **(0.202)**

Maximum likelihood estimates for the extremal index and three return levels for the Newlyn sea–surges (units for return levels are in metres).

- Suitability of first order Markov assumption/models used?
- Dependence of some estimators for θ on auxiliary parameter
- $\hat{\theta}^{[5]}$ most suitable?

- An extensive study in Fawcett & Walshaw (2006) suggests that a first–order Markov structure, with logistic dependence, is suitable
- So $\hat{\theta}^{[1]}_{\text{log}}$ might be a suitable estimator for the extremal index
- Need to account for seasonality: monthly varying GPD parameters can be combined to estimate overall return levels by solving

$$\prod_{m=1}^{12} \mathcal{H}_m(x)^{n_m \theta_m} = 1 - r^{-1}, \qquad m = 1, \ldots, 12.$$

3.2 Data applications: Bradfield wind speeds

	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
All excesses			
using $\hat{\theta}^{[1]}_{\log,m}$	88.463 (5.520)	96.071 (9.967)	107.644 (22.435)
using $\hat{\theta}^{[5]}_{m}$	84.885 (6.151)	92.882 (8.873)	105.003 (19.745)
Cluster peaks	96.556 (13.527)	102.537 (22.776)	107.143 (43.052)

Maximum likelihood estimates for three return levels for the Bradfield wind speeds (units are in knots).

3.3 Confidence intervals for return levels

- Standard errors highlight the gain in precision when using all threshold excesses

- We obtain confidence intervals using the **bootstrap distribution** for z_r
  - Use a block bootstrap procedure to sample, with replacement, entire clusters of extremes
  - For each bootstrap replication b, $b = 1, \ldots, B$, find $\left(\hat{\lambda}_u, \hat{\sigma}, \hat{\xi}, \hat{\theta}^{[5]}\right)^{(b)} \longrightarrow \hat{z}_r^{(b)}$
  - Form 95% confidence interval for z_r from the bootstrap sample $\left\{\hat{z}_r^{(1)}, \ldots, \hat{z}_r^{(B)}\right\}$
  - We use **bias–corrected, accelerated** intervals (Efron, 1987), which give better coverage than the standard **percentile** intervals

3.3 Confidence intervals for return levels

	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
Sea surges	$(0.657, 0.872)$	$(0.708, 1.019)$	$(0.772, 1.306)$
Cluster peaks	*(0.765, 1.569)*	*(0.792, 2.675)*	*(0.835, 6.452)*
Wind speeds	$(80.847, 87.749)$	$(86.088, 98.540)$	$(90.623, 116.103)$

Bootstrapped 95% (BC$_a$) confidence intervals for three return levels for the Newlyn sea–surges (metres) and the Bradfield wind speeds (knots).

Summary

- Return level inference under the standard POT approach can be highly sensitive to the choice of declustering interval used to identify clusters
- Using *all* threshold excesses can avoid the issue of declustering
  - This requires an appropriate estimator of the extremal index
  - The intervals estimator seems robust here
- Using *all* threshold excesses can substantially increase precision of return level estimates
- A block bootstrap procedure can be used to obtain confidence intervals

References

- Efron, B. (1987). Better Bootstrap Confidence Intervals. *J. Am. Stat. Ass.*, **82**(397), pp. 171—185.
- Fawcett, L. and Walshaw, D. (2012). Estimating return levels from serially dependent extremes. *Environmetrics*, **23**(3), pp. 272—283.
- Fawcett, L. and Walshaw, D. (2006). Markov Chain Models for Extreme Wind Speeds. *Environmetrics*, **17**(8), pp. 795—809.
- Leadbetter, M.R. and Rootzén, H. (1988). Extremal theory for stochastic processes. *Ann. Probab.*, **16**, pp. 431—476.
- Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Stat.*, **3**, pp. 119—131.